J Comput Neurosci (2008) 25:562–582
DOI 10.1007/s10827-008-0094-6

Path-finding in real and simulated rats: assessing the influence of path characteristics on navigation learning

Minija Tamosiunaite · James Ainge · Tomas Kulvicius · Bernd Porr · Paul Dudchenko · Florentin Wörgötter

Received: 25 June 2007 / Revised: 11 March 2008 / Accepted: 24 March 2008 / Published online: 30 April 2008

Abstract A large body of experimental evidence suggests that the hippocampal place field system is involved in reward based navigation learning in rodents. Reinforcement learning (RL) mechanisms have been used to model this, associating the *state space* in an RL-algorithm to the place-field map in a rat. The convergence properties of RL-algorithms are affected by the exploration patterns of the learner. Therefore, we first analyzed the path characteristics of freely exploring rats in a test arena. We found that straight path segments with mean length 23 cm up to a maximal length of 80 cm take up a significant proportion of the total paths. Thus, rat paths are biased as compared to random exploration. Next we designed a RL system that reproduces these specific path characteristics. Our model arena is covered by overlapping, probabilistically firing place fields (PF) of realistic size and coverage. Because convergence of RL-algorithms is also influenced by the state space characteristics, different PF-sizes and densities, leading to a different degree of overlap, were also investigated. The model rat learns finding a reward opposite to its starting point. We observed that the combination of biased straight exploration, overlapping coverage and probabilistic firing will strongly impair the convergence of learning. When the degree of randomness in the exploration is increased, convergence improves, but the distribution of straight path segments becomes unrealistic and paths become 'wiggly'. To mend this situation without affecting the path characteristic two additional mechanisms are implemented: A gradual drop of the learned weights (weight decay) and path length limitation, which prevents learning if the reward is not found after some expected time. Both mechanisms limit the memory of the system and thereby counteract effects of getting trapped on a wrong path. When using these strategies *individually* divergent cases get substantially reduced and for some parameter settings no divergence was found anymore at all. Using weight decay and path length limitation at the same time, convergence is not much improved but instead time to convergence increases as the memory limiting effect is getting too strong. The degree of improvement relies also on the size and degree of overlap (coverage density) in the place field system. The used combination of these two parameters leads to a trade-off between convergence and speed to convergence. Thus, this study suggests

Action Editor: Alain Destexhe

M. Tamosiunaite · T. Kulvicius
Department of Informatics, Vytautas Magnus University, Vileikos 8, 44404 Kaunas, Lithuania

M. Tamosiunaite
e-mail: m.tamosiunaite@if.vdu.lt

M. Tamosiunaite · J. Ainge · P. Dudchenko · F. Wörgötter
Department of Psychology, University of Stirling, Stirling FK9 4LA, Scotland

T. Kulvicius · F. Wörgötter (✉)
Bernstein Center of Computational Neuroscience, University Göttingen, Göttingen, Germany
e-mail: worgott@bccn-goettingen.de

T. Kulvicius
e-mail: tomas@bccn-goettingen.de

B. Porr
Department of Electronics & Electrical Engineering, University of Glasgow, Glasgow, GT12 8LT, Scotland
e-mail: B.Porr@elec.gla.ac.uk

that the role of the PF-system in navigation learning cannot be considered independently from the animals' exploration pattern.

Keywords Reinforcement learning · SARSA · Place field system · Function approximation · Weight decay

1 Introduction

The learning of goal-directed navigation in rats and other rodents is one likely function of their hippocampal place field system. Several studies exist that use variants of reinforcement learning (RL) algorithms to show that the place field system could indeed serve as a substrate for navigation learning (Foster et al. 2000; Arleo and Gerstner 2000; Arleo et al. 2004; Strösslin et al. 2005; Krichmar et al. 2005). In general, RL assumes that the space in which learning takes place is tiled into states where certain actions can be taken from every state to reach a goal (e.g. a reward location). Through exploration of the state space, an agent will try out different actions at different states and in this way it can recursively find the best possible route (the optimal policy) to a goal (for a review see Sutton and Barto 1998). In the most general case of a RL system, all states in the state space will have to be visited "often enough" to try out the different actions necessary for convergence. This can, however, lead to a problem because convergence is very slow if the combined state-action space is large (the "curse of dimensionality" problem). Thus, in big state spaces, value function approximation versions of RL algorithms are used (Tesauro 1995; Sutton and Barto 1998). These cover the state space with large, possibly overlapping kernels and run RL over this feature space, instead of iterating over every individual state.

A second problem concerns the way RL algorithms usually choose exploration strategies. In order to learn, the agent has to explore the state-action space. Proofs exist that sufficiently dense, unbiased exploration will lead to convergence to the optimal solution in the most common RL-algorithms (Sutton and Barto 1998). To this end, conventional RL methods use random exploration, which in a navigation task leads to random walk patterns that appear incompatible with biological paths. Animals typically produce more ballistic (straight) exploration paths with only a limited degree of randomness, the length of which gradually increases from a home-base into the unknown terrain. Their paths often follow walls and landmarks, especially in daylight (Etienne et al. 1996; Eilam 2004; Zadicario et al. 2005). This, however, leads to an exploration bias that jeopardizes the convergence of the RL methods.

Thus, in this study we focus on the interaction between path shapes and learning in a simulated rat. The place field representations we use are abstract. Thus, our intention is not to produce a model of the hippocampus and its function. Rather, this study will focus on the interaction between biological path generation strategies and the convergence properties of learning. Specifically, we will also explore how the extent of the overlap and coverage of the location representations, the place fields, affect convergence and the speed to convergence.

We implement different path generation strategies that are realistic in that they reproduce specific statistical properties of actual rat paths, recorded and analyzed for this study. We observe that convergence of RL is not generally assured when using a realistic exploration pattern by our simulated rats. We will, however, show that our system can be stabilized by weight decay or path length limitation. These two mechanisms are in the Discussion section linked to bio-psychological aspects of forgetting and frustration.

2 Methods

The study uses methods from RL with function approximation to achieve fast convergence. The description of these methods is quite technical. Hence we present the RL methods in the Appendix as it is not of central interest for the topics of this study. Here it may suffice to explain that in this study we are using on-policy SARSA learning. SARSA stands for "state-action-reward-state-action" referring to the transitions an agent goes through when learning (Sutton and Barto 1998). This is motivated by recent findings in the midbrain dopaminergic system (Morris et al. 2006). Alternatives would be Q-learning or Actor-Critic Learning and we will in detail discuss the choice of SARSA-learning in the Discussion section.

2.1 Model environment

Our model animal performs a simple navigation towards a goal task in a homogeneous environment similar to a Morris water maze task (Morris 1984), in that there are no odor cues or obvious landmarks. In Fig. 1(a) a schematic picture of the model environment is provided.

The model environment is 150 cm × 150 cm, discretized using a grid of 10000 × 10000 units. The model animal at each learning trial is placed at a predefined

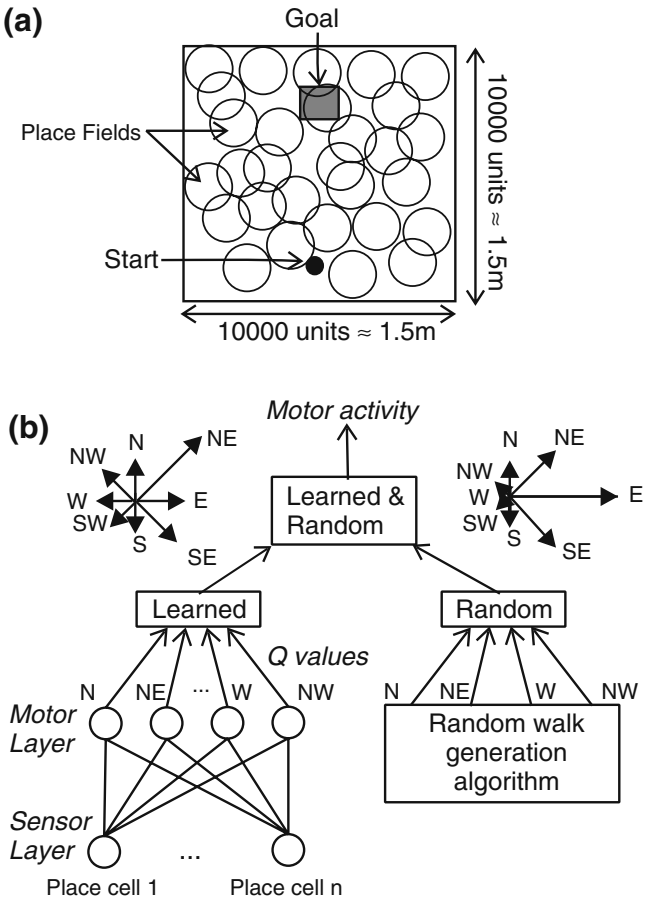

Fig. 1 Model environment: the start is shown as a *circle*, and the goal is a *rectangular area* of the size 15 cm × 15 cm, located opposite to the start 15 cm away from the upper border. *Big circles* schematically show place fields covering the arena. (**b**) Neural network of a model animal where motor activity is obtained as combination of learned direction values (Q-values, *Learned*) and path straightening components (*Random*). The *stars* show examples how these components could look like. The two components are combined summing learned and random components for each direction with appropriate coefficients (*Learned & Random*), and then choosing the prevailing direction for a motor action

position (75 cm, 15 cm). A reward of the size 15 cm × 15 cm is placed opposite to the start 15 cm away from the upper border of the environment. The model animal travels in predefined steps (6 cm ± a random component of up to 1.5 cm). After the model animal reaches the reward or does not reach the reward in a predefined number of steps the animal is reset to the start position for the next trial.

The substrate for learning in our system are simulated place fields distributed within an arena. We assume that a place cell i produces spikes with a scaled Gaussian probability distribution:

$$p(\delta_i) = A \exp\left(-\delta_i^2 / 2\sigma^2\right) \quad (1)$$

where δ_i is the distance from the i-th place field center to the sample point on the trajectory, σ defines the width of the place field, and A is a scaling factor. In the areas where the values of this scaled distribution are above 1, cells fire with a probability of 1.

Place cell centers are distributed in the model environment randomly, with a uniform distribution. Experiments are performed with 20-2000 cells. Field width is defined by $\sigma = 2.12, 4.24, 6.36, 8.48$ cm. Fields are cut wherever they touch a boundary. A scaling factor of $A = 2.5$ Eq. (1) is applied to the probability distributions of place cell firing, to make cell spiking more regular inside a place field. The size of a place field can be estimated by its firing probability cut off, e.g. 10%, using: $size_{P=0.1} = 2\sigma\sqrt{-2ln(P/A)}$. This yields, for $\sigma = 4.24$ cm, a firing field of about 22 cm size, which is in correspondence with the literature (Muller 1996; O'Keefe and Burgess 1996; Mehta et al. 1997).

2.2 General scheme for navigation

We are investigating learning in a network composed of two layers of cells (see Fig. 1(b)). At the lower layer of the network are the place cells. In the upper layer are motor cells, which learn to perform the navigation task. To keep the setup simple, we do not model head direction cells that are often also included in hippocampus-like navigation models (Brown and Sharp 1995; Arleo and Gerstner 2000), but allow the motor cells to direct the model animal movement towards eight directions: North, North-East, East, South-East, South, South-West, West and North-West. The actual direction is obtained combining motor cell outputs and path straightening components, used for a realistic path forming strategy (see below and also Fig. 1(b)).

2.3 Path generation and exploration strategies

In general, if the model animal did not attain the goal in 300 steps, it was reset to the start position for a new trial. This applies to all strategies (E,S,F,L), which will now be introduced one by one. Two path forming strategies are employed.

E-strategy is a usual RL strategy, with exploration and exploitation, where the path is chosen according to the learned Q-values most times, (probability $1 - p_e$), and a random move is made with probability p_e, where $0.1 \leq p_e \leq 0.2$. For random moves all directions are given equal probability. If not stated otherwise, we have set $p_e = 0.2$.

S-strategy performs straightening of the paths, where probabilities $p_1, p_2, ..., p_8, \sum_i p_i = 1$, are used, depending on the direction of the previous step. We define p_1

as the probability to proceed along the same direction as before, p_2 and p_8 then correspond to 45° to the left and right of this direction, p_3, p_7 reflect 90°, etc. In most of our studies we exclude backwards movement, setting $p_5 = 0$, to prevent an animal from performing small forward-backward cycles. When Q values are present, a weighted mixture of Q-based drives and randomized path-straightening drives is used:

$$d_1 = wq_1^N + (1-w)p_1$$
$$d_2 = ...$$
$$...$$
$$d_8 = wq_8^N + (1-w)p_8 \quad (2)$$

where $d_1, ..., d_8$ are the final drives, $q_1^N - q_8^N$, the normalized Q-values of the eight possible directions. Normalization is used to get $\sum_i q_i^N = 1$, and to bring Q-values in correspondence with the probabilities of the randomized drives. As default, we have used $p_1 = 0.5$, $p_2 = 0.156$, $p_3 = 0.063$, $p_4 = 0.031$, $p_5 = 0$, $p_6 = 0.031$, $p_7 = 0.063$, $p_8 = 0.156$, $w = 0.5$.

We also investigated a mixture of strategies E and S, where Q-values with straightening *and* some random exploration $0.1 \leq p_e \leq 0.2$ were used.

2.4 Weight decay and path length limitation

We are dealing with a learning system based on function approximation using place fields, which samples the space in a biased way (S-strategy). As a consequence this system can get trapped in divergent paths. Against this weight decay should help as the animal gradually forgets wrong paths, while path length limitation reduces the danger of trapping to begin with. Path length limitation can be linked to the return to home base behavior found in rats (Eilam and Golani 1989; Whishaw et al. 2001; Wallace et al. 2002; Hines and Whishaw 2005; Nemati and Whishaw 2007). There are other psychologically motivated variables which can influence learning, like surprise, hunger, mood, fatigue, etc. In the context of this study these variables were not modeled as they do not match to the time scale of individual rat trials. Weight decay and path length limitations act on every single trial, surprise has a shorter time scale (acting at one moment in time), while the other variables act on longer time scales (across many trials). Hence, introducing other variable would make the model at this state unduly complex.

Weight decay (*F-strategy, F for forgetting*) is implemented with a slow exponential decay characteristics:

$$\theta(t+1) = c_f \theta(t) \quad (3)$$

where c_f is in the interval 1.0-0.99, and 1.0 represents no weight decay. Each θ is a learned weight between a place cell and a motor cell, coding for the usefulness of moving into the direction represented by the motor cell when the rat is at the position represented by the place cell. A formal description for θ is given in the Appendix. When in use, the F-strategy is applied to all weights from sensor to motor layer in each step of a model animal. If not stated otherwise we use $c_f = 0.9995$ for the experiments. If weights fall below a threshold t_f due to decay, they are set to zero. We used $t_f = 0.000001$. Note, as weight decay happens step by step these apparently small decay rates act in an exponential way and decay over long paths becomes indeed quite strong.

Path length limitation (*L-strategy*) is implemented as a return to the start position if the reward is not found within an expected number of steps. Hence in this case the trial is aborted. Initially we limit learning to $k_l = 200$ steps. If the reward is found in trial n within $k(n) \leq 200$ steps, we set the limit k_l for the maximally allowed number of steps for the next trial $n+1$ to $k_l(n+1) = k(n) + \sqrt{k(n)}$. From there on, for every occurring failure trial, where the reward has not been found within the currently allowed path length limit, we increase the limit $k_l(n+j)$, $j \geq 2$ by a constant c_l using:

$$k_l(n+j) = k(n+1) + (j-1)c_l, \quad j \geq 2 \quad (4)$$

If the reward is then again found in trial m within the currently allowed limit we reset the limit k_l to $k_l(m+1) = k(m) + \sqrt{k(m)}$ and the counter to $j = 1$.

Thus, exceeding the limit for the first time leads to a Weber-Law like increase, where we use the square root function instead of the logarithm for simplicity. For every following trial, where the reward is not found, k_l is increased by some constant c_l, which resembles a relaxation process that gradually widens the exploration horizon. There is no rigorous data about this type of behavior in real animals, but it is known that rats start to explore again more and more, possibly due to gradually increasing hunger and/or motivation. Furthermore we note that during unsuccessful trials nothing is learned and Q-values are not updated. This is motivated by the situation that a real animal can only learn when the reward is indeed found.

Parameters of the model system are provided in a condensed way in Table 1.

2.5 Experimental methods

A total of 5 male Lister hooded rats weighing between 300-400 g were used. In this and the subsequent

Table 1 Default parameters used for modeling experiments in a standard setup

Parameter type	Parameter name	Value
SARSA-learning (see Appendix)	Learning rate α	0.7
	Discount factor γ	0.7
Environment/steps	Size	150 cm × 150 cm
	Step size	6 cm
	Noise on the step size	± 1.5 cm
	Reward size	15 cm × 15 cm
Place fields[a]	Number	500
	Width, through σ	4.24 cm
	Scaling factor A	2.5
Learning strategies	Exploration probability p_e in E	0.2
	Probabilities for S	
	p_1	0.5
	p_2	0.156
	p_3	0.063
	p_4	0.031
	p_5	0
	p_6	0.031
	p_7	0.063
	p_8	0.156
	Weighting factor w in S	0.5
	Weight decay factor c_f in F	0.9995
	Zero weight threshold t_f in F	10^{-6}
	Starting path length limitation in L	200
	Path increase step in L, c_l	5
	Path limit in steps for any strategy	300

[a]Note, additional justification for these default parameters is given in section 'Place field size and density'

experiments, compliance was ensured with national (Animals [Scientific Procedures] Act, 1986) and international (European Communities Council Directive of 24 November 1986 [86/609/EEC]) legislation governing the maintenance of laboratory animals and their use in scientific experiments. The rats were equipped with chronic recording electrodes, as described by Ainge et al. (2007), although the primary interest of this experiment was the behavior of the rats in finding a goal location. Rat trials were recorded in a square shaped arena of size 1.5 × 1.5 × 0.4 (length, width, height in meters) with blue walls. Each wall was equipped with small (10 × 10 cm) black felt "curtains", spaced equally from one another along the base of the wall. At the beginning of each trial the rat was placed at the same start location close to the center of the arena. A small piece of food (a chocolate cereal loop) was presented to the rat by the experimenter whenever the rat approached a pre-specified curtain. Rats were initially unfamiliar with this arena, and 10-40 trials were run with the rat being rewarded with food whenever it approached the "correct" curtain. The total number of trials depended on the rat's motivation and learning performance. Our measure of performance was the directness of the rats' paths to the correct location. After obtaining a reward, the rat was put back into a smaller opaque container (50×50 cm) for a short inter-trial interval. We have also manually removed the rat from the arena when it stopped searching for food, because these animals do not have a real home base to which they could run back, which would be their normal type of behavior in such a case. These cases are, however, rare for a motivated (hungry) rat and do not influence the path statistics. The position of the rat was monitored during the recording session through a black and white camera mounted on the ceiling above the arena. Two groups of ultra-bright LEDs were attached to the end of the recording cable, which in turn was connected to the chronic electrode. The LEDs were tracked using a recording system (Axona Ltd., St. Albans, UK), which detected the position of the two lights, thus providing information regarding the rat's location and the direction that the rat was facing at a sampling rate of 50 Hz. Data from the LED coordinates were stored on the hard drive of a PC.

2.6 Analysis of path statistics

For real and simulated rat trials, we determined the length of straight path segments and also how often they turn (directional change). Straight segments have been determined by standard linear regression moving

Table 2 Tables for the Kolmogorov-Smirnov test of real versus simulated rat segment length (top) and turning angle (bottom) distributions

Segments							Angles					
Start, average n=1494			Middle, average n=358		End, average n=391		Start, average n=7357		Middle, average n=2389		End, average n=3110	
Rats vs.	D	η	D	η	D	η	D	η	D	η	D	η
S	0.0402	0.0903	0.2470	0.3868	**0.2896**†	0.2624	0.0430	0.0441	0.0512	0.1367	0.1074	0.1188
SE	0.0744	0.0973	0.1389	0.2883	0.2017	0.2519	0.0229	0.0442	0.0704	0.0850	0.0972	0.1130
SL	0.0595	0.1101	0.0796	0.1866	0.2464	0.2550	0.0446	0.0485	0.0472	0.0810	0.0989	0.1148
SF	0.0705	0.0971	0.0872	0.2150	0.2041	0.2558	0.0369	0.0437	0.0509	0.0902	0.0736	0.1162
SEF	0.0793	0.0964	0.0840	0.1970	0.2456	0.2563	0.0229	0.0439	0.0790	0.0845	0.0646	0.1148
SEL	0.1035	0.1103	0.1088	0.1798	0.1769	0.2537	0.0222	0.0494	0.0730	0.0794	0.0700	0.1147
SLF	0.0448	0.1101	0.0512	0.1830	0.2005	0.2573	0.0472	0.0483	0.0455	0.0799	0.0972	0.1157
SELF	0.0823	0.1086	0.1059	0.1895	0.2480	0.2539	0.0210	0.0484	0.0720	0.0828	0.0667	0.1143
E	**0.3441**	0.0915	**0.3221**	0.1848	**0.4379**	0.2615	**0.2357**	0.0432	**0.1252**	0.1188	**0.1208**	0.1155
EL	**0.3603**	0.0931	**0.3616**	0.1783	**0.3270**	0.2603	**0.2427**	0.0440	**0.2514**	0.0801	**0.2166**	0.1197
EF	**0.3527**	0.0909	**0.2125**	0.2076	0.2486†	0.2622	**0.2303**	0.0429	**0.2088**	0.0902	0.1056†	0.1169
ELF	**0.3561**	0.0948	**0.3153**	0.1753	**0.2677**	0.2521	**0.2294**	0.0448	**0.2249**	0.0783	**0.1628**	0.1147

"Start" refers to the first, "middle" to the second and "end" to the last 1/3rd of all trials in an experiment excluding some trials as described in the text. Sample sizes n are in general large but can vary a lot and only averages are given. Thresholds η for the 1% significance level are given as well as the test variable D. Numbers in bold-face indicate that the Null hypothesis of similarity of the distributions must be rejected at the 1% level. The dagger symbols mark three cases which do not match with our prior expectations (see text)

along the path on a sliding window. This window was extended along the signal until a threshold for the average residual of $r \approx 1.25\%$ was reached corresponding to 2 cm. For real rats, we reset the analysis window with every stop. The choice of this threshold will clearly influence the placing of the break-points between segments, and local analysis with a sliding window will not give the optimal division of a path into longest possible straight segments, but we are not concerned with an exhaustive analysis. We are only interested in generating paths with statistical properties that are similar enough to the real rat trials. Using the same threshold relative to the environment size for real and simulated trials will allow for this comparison. Using this algorithm, we computed the segment length distributions for real and simulated rats for different cases as shown in the results section.

Furthermore, we calculated the turning angle distributions. For this, we move along the path (real and artificial) in predefined steps, such that the step length takes the same proportion of the arena both in the real and artificial example, and evaluate the angle between each two successive steps. We then bin angles into 8 categories: zero degrees turn, ±45 degrees turn, ±90 degrees turn, ±135 degrees turn, and 180 degrees turn to arrive at a distribution.

For a quantitative comparison between real and simulated rat path distributions we are using the Kolmogorov-Smirnov test (Stuart et al. 1999). The Kolmogorov-Smirnov test is known as the sharpest statistical test for comparing two different (unknown) distributions because it is sensitive not only to mean and median changes but also to skewness and kurtosis. We are testing the distributions against the threshold η for a 1.0% significance level, which is a strong criterion. Note, the threshold depends on sample sizes n of both distributions. Tested is the Null-hypothesis that two distributions are identical by comparing test variable D against the threshold. If $D > \eta$ then the Null-hypothesis needs to be rejected and distributions are different at the 1.0% significance level. (see Table 2).

3 Results

3.1 Qualitative analysis of real and simulated rat paths before learning

Figure 2 shows examples of real and simulated rat trials and their statistical properties. As we are first concerned with setting up the initial conditions for our model in an appropriate way, we will now describe path characteristics of real and simulated rats without learning focusing on visual inspection of the presented data.

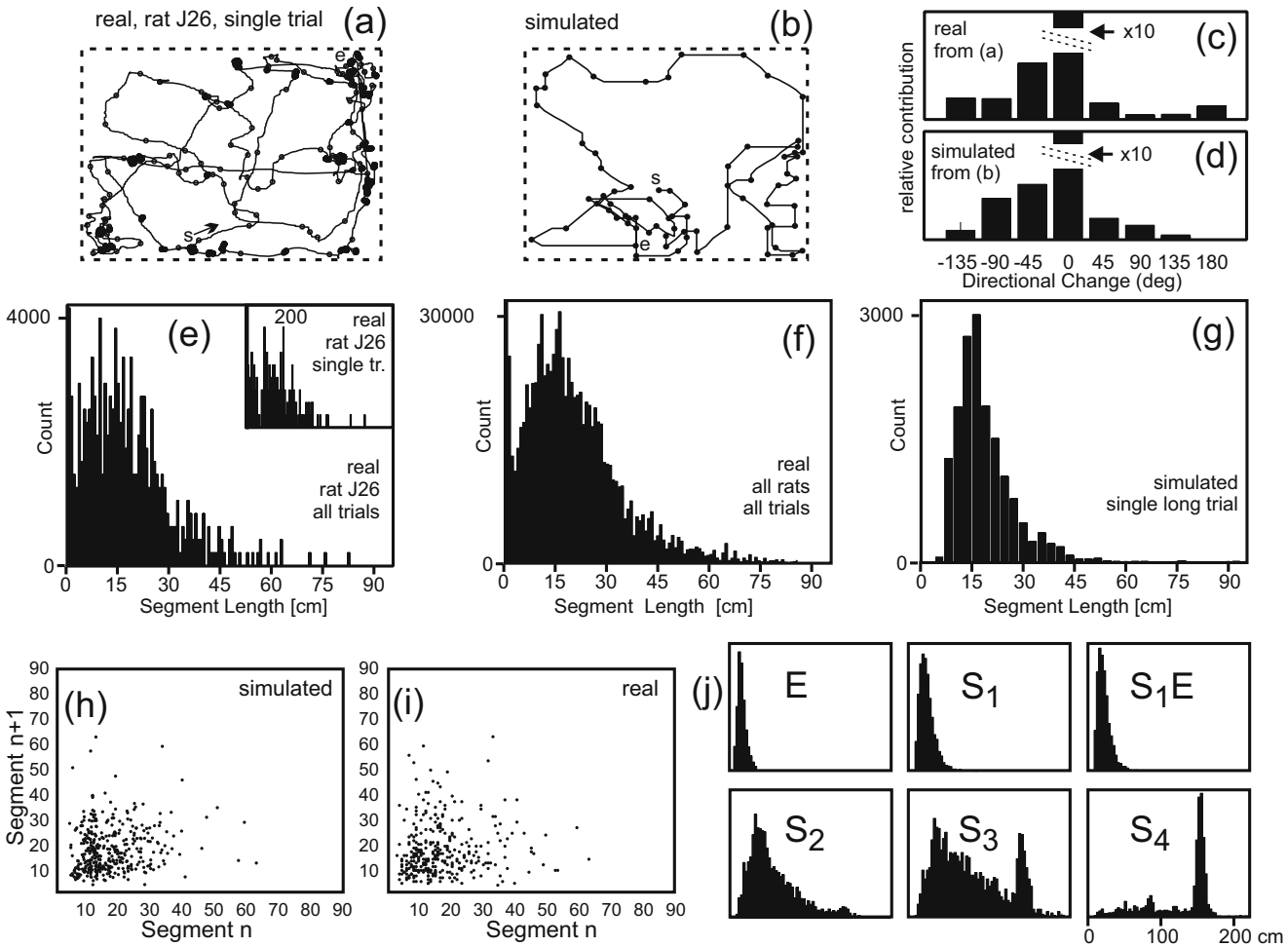

Fig. 2 Statistical analysis of real and simulated rat trials. (**a**) Example of a real rat trial in a rectangular arena. Walls of the arena are shown by the *dashed box*. *Dots* mark break points between straight stretches. (**b**) Simulated trial. Start and end-points are marked (*s*, *e*). (**c**, **d**) Distribution of turning angles for a real (**c**) and a simulated (**d**) trial. (**e**–**g**) Distribution of straight stretches for real (**e**, **f**) and simulated (**g**) trials scaled to their individual peak height. Panel (**e**) shows the distribution of straight stretches for a single rat (all trials), inset in (**e**) shows a single rat trial as given in panel (**a**). Panel (**f**) contains all five experimental sessions with real rats and (**g**) simulated trial. The simulated path was generated using 'SE' strategy (see subsection 'Path Generation and Exploration Strategies'). Panels (**h**) and (**i**) show scatter plots of subsequent path segments. Small panels (**j**) depict the distributions from trials with different S and E components as shown by the labels. For S1 the default parameters were used, for S2, S3 and S4 we have S2: $p1 = 0.8475$, $p2 = 0.0656$, $p3 = 0.0087$, $p4 = 0.0019$, $p5 = 0$, $p6 = 0.0019$, $p7 = 0.0087$, $p8 = 0.0656$; S3: $p1 = 0.9386$, $p2 = 0.0286$, $p3 = 0.0018$, $p4 = 0.0002$, $p5 = 0$, $p6 = 0.0002$, $p7 = 0.0018$, $p8 = 0.0286$; S4: $p1 = 0.9940$, $p2 = 0.0030$, $p3 = p4 = p5 = p6 = p7 = 0$, $p8 = 0.0030$

Only after having introduced the different learning properties, we will compare simulated with real paths also during learning, providing also a large quantitative comparison based on statistical distributions (see Table 2).

The example of a real rat path in (a) shows that rats have the tendency to continue on their path for some time often along the walls and exploring inwards. Smooth curves are rare; instead the animals turn rather sharply. Accumulations of dots occur at locations where the rat had stopped and performed behaviors such as sniffing or resting. Panels (e, f) show segment length distributions from several real rat trials. Stopping and sniffing creates almost all the contributions for the leftmost bins in the distributions (e, f), where we have clipped the bin for segment length 1.0 because a huge number of such mini-segments occur when a rat stops and just moves its head. In general, the distributions for single trials (inset in e), all trials of one rat (e), and all trials of all rats (f) are smoother as the amount of data increases.

In (b) we show a simulated trial that has been generated using the SE strategy. Judging by eye, real (a) and simulated paths (b) appear similar. Also, the segment

length distribution of one simulated long trial (g) is similar to (f), very small segments, however, occur very rarely in the simulations as simulated rats do not stop.

Averaging over more simulated trials will lead to a smoother distribution (not shown), but does not otherwise alter its shape as the same generative algorithm is always used. Panels (h) and (i) present scatter plot of subsequent path segments in simulated (h) and real (i) rat trials. Also here there are no clear differences visible.

Distributions (c) and (d) represent the number of turns and their degree for a given path. We have binned angles into 8 categories: zero degrees turn, ±45 degrees turn, ±90 degrees turn, ±135 degrees turn, and 180 degrees turn. Note, the zero-bin is 10 times larger than shown in the histograms. Both distributions are similar, and somewhat skewed to the left as the actual rat trial used was dominated by a leftward running tendency (see a, b). The 180 deg bin is empty in the simulated trials, because we did not allow the rat to directly run back. For the real rat some entries in this bin are probably due to switchbacks that occurred while stopping.

Finally the small panels (j) at the bottom; labeled E, S_1, S_1E, S_2, S_3, S_4; depict different distributions of simulated trials. Specific parameters for path generation are given in the figure legend. Note scaling of the x-axis is here 210 cm and not 90 cm as above. Panel S_1E shows the same case as (g). Panel E contains no path straightening and, as a consequence, small segments begin to dominate. S_1, on the other hand, contains only the S component. In spite of missing E, it is still very similar to S_1E. Hence, concerning path characteristics cases S_1E and S_1 are essentially the same as will be quantified in Table 2, below. Cases S_2, S_3 and S_4 show what happens when we change the asymmetry in the path generation algorithm toward increasingly straighter path segments. S_4 is an extreme case, where the rat runs almost all the time along the walls. Very long straight segments dominate in this case and this distribution is strongly different from any of the ones above.

3.2 Artificially generated path shapes with and without learning and quantitative comparison to real rat trials

The previous section showed data and a qualitative comparison of real and artificial paths before learning. Now we will show examples and provide a more detailed quantitative analysis of paths generated during learning. Figure 3 displays several more artificially generated paths to the reward. Examples are shown from early (panels a-d) as well as late (panels e-h) learning.

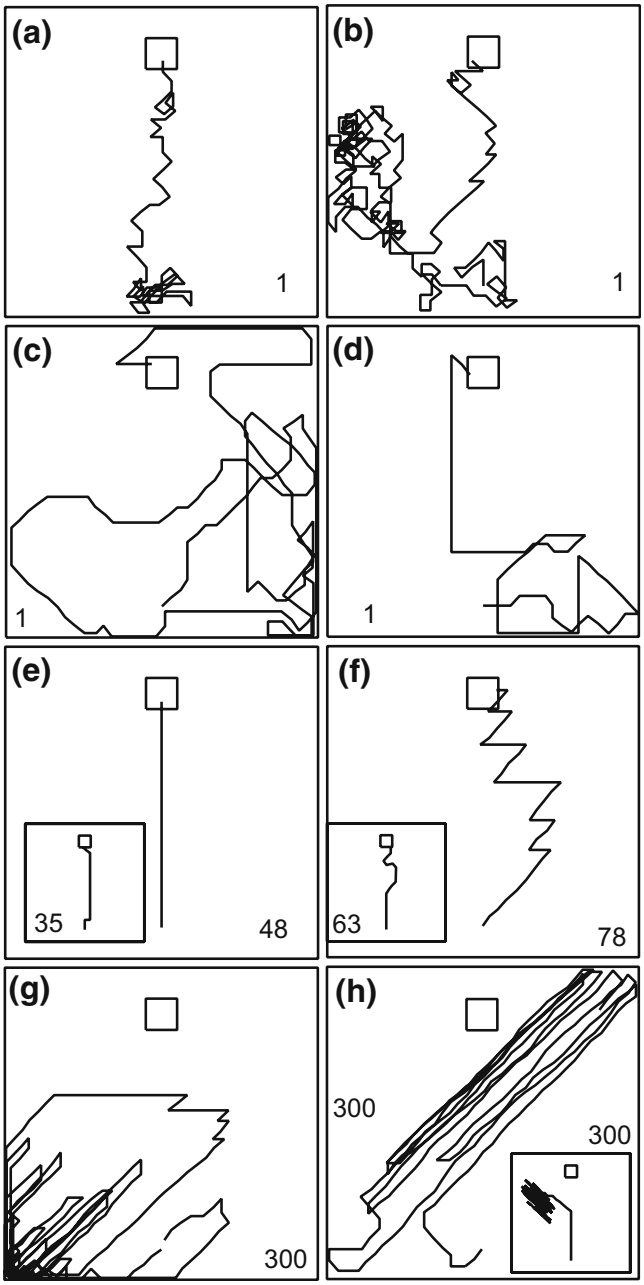

Fig. 3 Examples of paths obtained with SARSA learning under different strategies: (**a**, **b**) - traditional exploration-exploitation (E) for initial learning stages, (**c**, **d**) - exploration-exploitation mixed with path straightening (SE) for initial learning stages, (**e**) - learned optimal path with (S). When adding an E-component (SE) sometimes *kinks* exist from an exploratory move early on the path (inset in **e**), (**f**)- zigzagging learned path in straightened case (S). The *inset* shows that adding the E component (SE) will reduce zigzagging. Panels (**g**, **h**) show divergent paths for the straightened case (S), inset in (**h**) - divergent pattern when direction "back" is not forbidden. Default parameters (Table 1) were used. *Small numbers* at the bottom refer to the trial number from which the examples were taken

If paths are generated in the 'traditional' way by mixing exploration and exploitation (strategy E), as necessary to assure convergence of Q- or SARSA-learning, (panels a and b) they are 'wiggly' and do not resemble those of real animals. Similar to Fig. 2(a), mixing traditional E (exploration-exploitation) with path straightening (S) provides more realistic paths (Fig. 3(c, d)) when learning starts.

Figure 3(e) shows two optimal paths after learning. Using path-straightening strategy (S) only (or in conjunction with weight decay and path length limitation), straight paths were often learned. If an exploration component was added (mixing strategies E and S) a few off-path moves occur (inset in panel e). The learned components, however, assure that in these cases the rat gets back on track immediately. When using the strategy S, we find that SARSA learning can produce zigzagging paths at the end of learning (Fig. 3(f)). Note, when using Q-learning, such zigzagging does only occur in very rare cases, and curved paths occur instead. Zigzagging paths are normally not fully constant but zigzags will change often to a small degree trial by trial due to the random component in place cell firing which, as a consequence, also leads to some oscillations of the Q-values. Note, when using a regular (no variability) place field structure we are approximating an ideal SARSA learner and no more zigzagging occurs. When a mixture of strategies S and E is used, zigzagging is greatly reduced and - if convergent - paths are essentially stable in the end (inset in panel f). In such a situation nothing will change anymore, as we did not model motivational variables (like hunger), which would at some point again lead to more exploration in a real rat as soon as it is well fed. Panels (g) and (h) show divergent paths, which frequently develop when using strategy S. The inset in Fig. 3(h) shows a divergent pattern when the direction "back" was not forbidden, which for divergent cases often leads to very fast switchbacks. Divergent paths are often characterized by a weight vector field which points towards a boundary or corner such that the rat has little chance to escape and produces random loops like in (g).

In one of the following chapters we will compare the convergence properties of 12 different combinations of path formation and learning strategies. For this it is first necessary to show to what degree these different strategies produce realistic paths. To this end we used real and simulated trials to calculate segment and turning angle distributions (compare to Fig. 2) during learning. trials were subdivided into three learning phases (start, middle, end) by using the first, second and last $1/3rd$ of all trials from every experiment in a real or simulated rat. In simulated rats, for "middle" we excluded all trials where the target had not been found between 50 and 100 steps this way including only trials of medium length in the statistics. For "end" we excluded trials with more than 50 steps, to assure that only converged paths were included in the statistics. For real rats a similar procedure was adopted, calibrated against the minimal possible path length between starting point and reward. Note, "middle" trials are in general rare as real rats learn fast and keep running quite straight to the target as long as they are hungry, while afterward they begin to explore again which again leads to long trials.

We used the Kolmogorov-Smirnov test at the threshold η for the 1% significance level and tested the Null-hypothesis that two distributions are identical, which holds if $D \leq \eta$, where D is the test variable calculated by the test. Note, D and η depend on sample size and vary accordingly.

Table 2 shows that in all 24 cases, except 2, distributions of segments and angles from rat behavior differ from simulated behavior at a 1% significance level if the S-component is missing (cases E,EL,EF,ELF, bottom). With an S-component the situation is different and in all 48 cases, except 1, distributions are not significantly different at the 1% level. The three exceptions happen in the "end" learning phase. Here, simulated and real rats begin to run briskly towards target generating only a few path segments in each trial and distributions become rather featureless. This introduces a higher variance leading to the three outliers.

This allows the conclusion that simulated trials with an S-component are realistic with respect to segment length and angle distribution, while trials where the S-component is missing produce unrealistic statistics. Paths with random exploration E only, however, are often optimally convergent to the straight path between start and goal. Path straightening S, on the other hand, leads to zigzagging or even divergent paths. In the next section we substantiate these statements and try to provide a solution for combining realistic paths with good convergence.

3.3 Convergence patterns

First we consider individual examples of convergence for different combinations of path generation strategies (Fig. 4, number of steps to goal is shown). In the top row convergent cases for different path strategies are shown. In the bottom row cases, where convergence was less clear, are displayed. In panel (a) quick convergence to an optimal path is shown for a case of path straightening S. By optimal path we mean the one path which is straight between start and goal. The

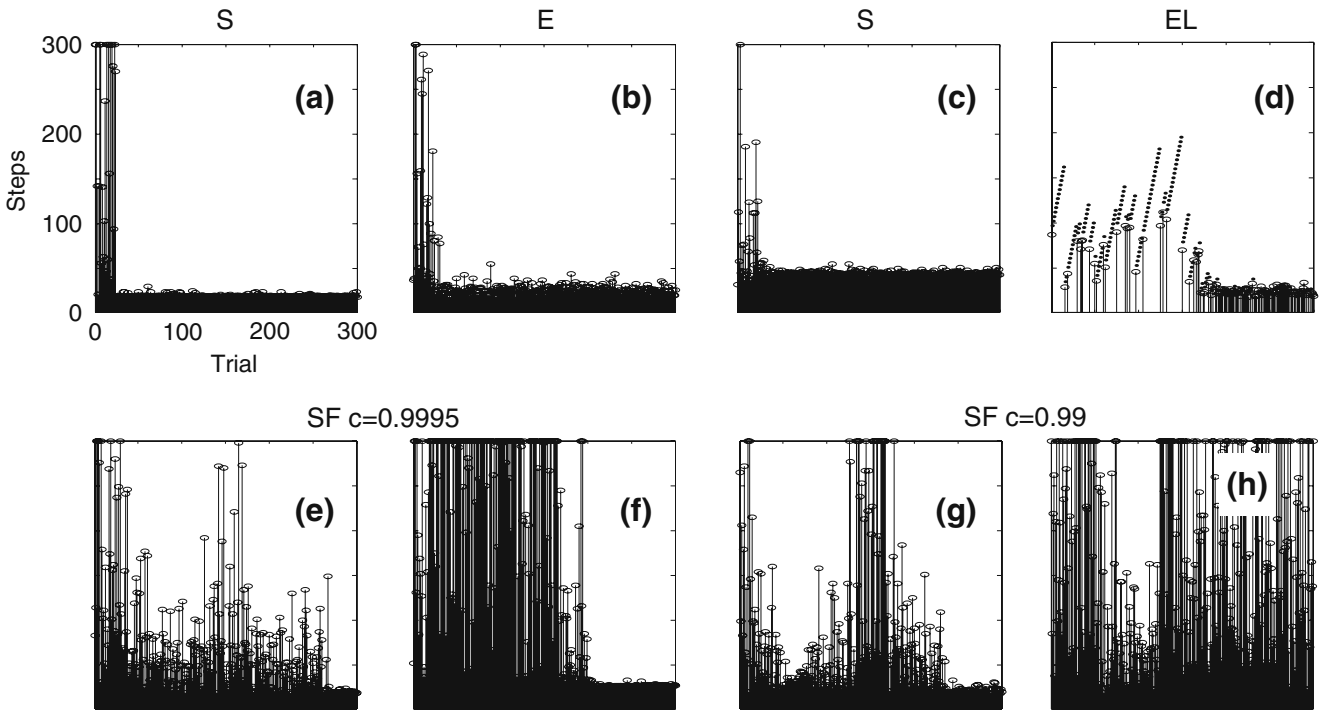

Fig. 4 Patterns of convergence for different path generation and learning strategies as given above each panel (*S*, *E*, *EL*, *SF*). (**a**, **b**) show situations where an optimal and (**c**) where a non-optimal (zigzagging) path has been learned. The *small dots* in panel (**d**) indicate trials where the rat had not found the reward within its limited learning horizon (strategy EL). Finally the optimal path has been found here, too. Panels (**e**–**h**) show different cases that can happen when using the SF-strategy. In (**e**) a very late convergence to the optimal path is seen. In (**f**) late convergence to a non-optimal path occurred. In (**g**) the system was intermediately divergent and finally found the optimal path, while in (**h**) the system seemed to converge, but then finally diverged. Default parameters (Table 1) were used in all cases

randomness in the spiking of the place fields leads to path lengths which vary minimally. In (b) quick convergence using traditional exploration-exploitation E is presented. One can see that here the path length fluctuates more due to occasional off-path steps (compare with inset in panel f of Fig. 3). In panel (c) a case is shown where the path straightening strategy S converges to a non-optimal (zigzagging) path-type and in panel (d) we give an example for convergence with limited path length L combined with exploration E. The dots indicate trials where the reward has not been found, which leads to a gradually growing path length limit. As a consequence of the L-mechanisms, it often happens that several successive trials are unsuccessful which leads to the upward slanted dotted "lines" in the diagram as the path length limit gets larger with every unsuccessful trial. At the end of these learning trials convergence to an optimal path is reached.

In cases (e–h) we show examples of convergence patterns for cases where we used straightened paths S with weight decay F. A longer time to convergence is a typical feature here and cases exist (e, g) where the behavior intermittently diverges. If weight decay is too strong, convergence is bad and even cases that begin to converge will at the end not find a good path (h).

Note, time to convergence (hence, the number of trials to find an optimal path to the reward) cannot easily be compared to real animals, because this depends again on the relative size of reward and arena. When being trained to find food, real rats found good paths in about 10 trials, which roughly compares to the simulation results shown here. Many times, however, after having learned, real animals would "get distracted" and kept on exploring, eventually going towards the reward much later. Furthermore, real animals also use other cues (like odor) for navigation (Save et al. 2000), which have not been modeled. Real animals sometimes show very strong inter-individual differences probably driven by more general states of, e.g., motivation, intention, fright, etc. Modeling behavioral states is beyond the scope of this study.

3.4 Statistical analysis of mixed strategies

In Fig. 5 a summary of the influence of the different mechanisms is presented. For each diagram 100

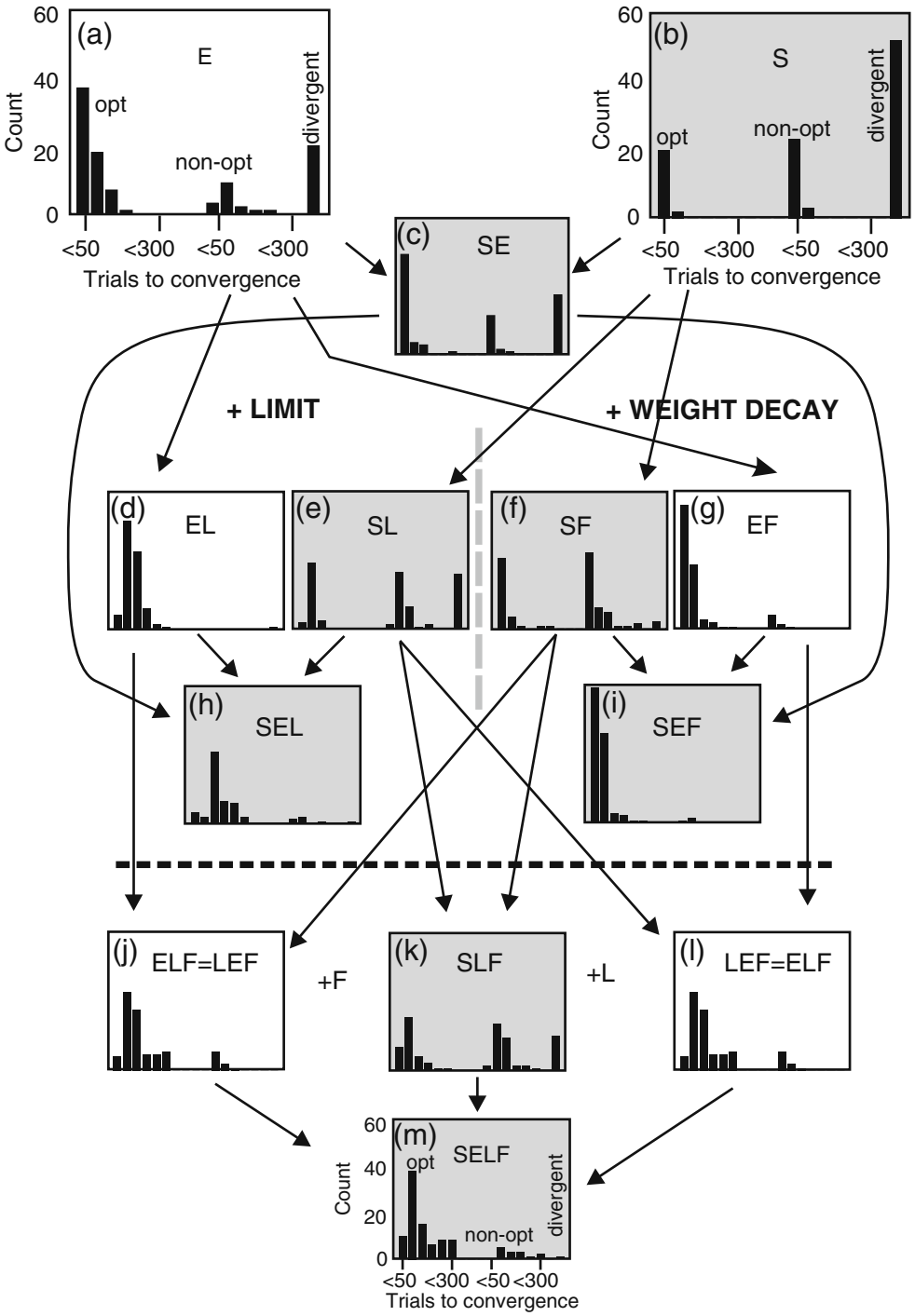

Fig. 5 Statistical analysis of mixed strategies for path generation and learning as given inside each panel for 100 experiments each. Arrows indicate how properties get inherited from each other. *Gray panels* are the ones containing an S component. Three histograms are shown in each panel: convergence to an optimal path (*left histogram*), to a non-optimal path (*middle histogram*) and divergence cases (*right single-bin*). For panels below the *dashed lines* the mixing of both memory limiting strategies - F and L - leads to a visible deterioration of the performance. Note, panels (*j*) and (*l*) are identical. Default parameters (Table 1) were used

experiments were analyzed. Arrows indicate how histograms inherit properties from top to bottom. The two top diagrams (a) and (b) show the basic cases of pure exploration E and pure path straightening S respectively. Below, cases with path length limitation L are shown (mostly) in the left and middle part of the diagram, cases with weight decay F are found right and middle. Towards the bottom more and more strategies are mixed and the left-right separation in the diagram vanishes.

Histograms in the figure show three groups of bins. In the leftmost group the optimal path has been found, the middle group shows cases where a sub-optimal path was found (e.g. zigzag) and the rightmost group, which consists of a single bin, shows the number of divergent cases. Bins in the groups are ordered to show cases of

convergence in less than 50 trials (leftmost bin) up to less than 300 trials (rightmost bin).

Gray shading indicates those cases where the paths contain an appropriate S component. This particular S-component leads to the situation in the main part of Fig. 2, which appears realistic relative to the other combinations shown there. Cases without shading produce unrealistic paths.

As expected, pure exploration (a) often leads to convergence, but several cases (20 out of 100) are observed where the system diverges. While biologically more realistic, the type of place field-like function approximation used here does not belong to the few known classes of function approximation algorithms for which convergence has been proved. With path straightening (b), the convergent cases converge faster (mean opt for S = 31.5 trials, for E = 56.9 trials), but there are many more divergent trials (52) now. "Mean opt" gives the mean value for the optimally convergent histogram (leftmost histogram). Mixing cases S and E (c) produces a result in between the pure S and E cases (mean opt for SE = 39.9 trials, divergent cases = 28).

Note, diagrams below this level (below panel c), which descend from an ancestor above (arrow) can be best understood by a leftward redistribution of the members in the bins of the respective ancestor, hence, one finds an improvement when going down. This picture generally holds well for all panels until (i), hence, above the dashed line, below of which too many mechanisms mix and performance deteriorates again. This will now be quantified in the following.

Limiting the learning horizon (L, panel d) or adding weight decay (F, panel g) efficiently eliminates all divergence from the pure exploration case (compare to a). As expected, learning is now slower though (mean opt for EL = 102.0 trials, for EF = 50.0 trials). If not concerned with realistic paths, strategies EF would be the best choice for fairly fast and robust convergence. Doing the same with straight paths (panels e, f) also leads to substantial improvement with respect to removing divergence cases as compared to (b) (divergent cases for SL=25, for SF=5), but convergence is again slower as in (b) (mean opt for SL = 76.4 trials, for SF = 38.4 trials).

In general this confirms the motivation for F and L presented in the Methods section as both mechanisms reduce the danger of getting trapped in a divergent situation.

Panels (j) and (l) represent the case (note, j and l are identical!) where the other (F or L) component has been added to cases EL and EF, respectively. Now learning becomes again slower because L and F *both* limit the memory of the learning system (mean opt = 126.4 trials, no divergence).

Mixing path straightening with a bit of exploration in general seems to be a good strategy (panels c, h, i, k, m), by which convergence is most of the time assured together with realistic looking paths. Case SEF (panel i) leads to fairly fast *and* robust convergence (mean opt for SEF = 45.6 trials, no divergence). Many cases were found where convergence happened within 10 to 30 trials, not much slower than in real rats. For case SEL (panel h) convergence was much slower (mean opt = 138.4 trials, one divergent case) and this also holds true when mixing both limiting strategies L *and* F in cases SLF (k) and SELF (m), where also more divergent cases begin to appear for SFL (18). Convergence times for SEL, SLF and SELF became, however, unrealistically long with a mean opt larger than 100 trials in all these cases.

In summary, when using realistic paths that contain an appropriate S-component (as judged by Fig. 2), convergence deteriorates and this suggests that other mechanisms are needed to counteract this effect, where here we used F and L. We find that straight paths together with weight decay and/or path length limitation will not lead to good performance (panels: e,f,k; representing cases SL, SF, SLF). Adding exploration (SE, c) will immediately improve on this, while still leaving the path shape realistic, but many divergent cases remain. This can be mended by adding weight decay (SEF, i), which represents the most realistic case concerning path shapes *and* convergence times. Path length limitation is also a powerful mechanism to eliminate divergence, but in these simulations convergence times became now rather long (h). Mixing too many strategies will also lead to performance deterioration, because they all work in the same way, reducing the memory of the system.

Furthermore, we investigated how sensitive the system reacts to parameters L and F, because, clearly, too much weight decay or path length limitation is also harmful. In Fig. 6 we show how the number of divergent cases depends on the decay rate for the SF case. It shows a minimum between 0.999 and 0.99999, and any weight decay rate within this range may be advantageously used. We also used path length limitation (SL) with constants $c_l = 2$, $c_l = 3$, $c_l = 5$, and $c_l = 7$, and found that with $c_l = 2$ the process converges about 60% slower as compared to the here used standard case of $c_l = 5$ (mean opt for $SL_{c_l=2} = 124.3$ steps as compared to mean opt for $SL_{c_l=5} = 76.4$ steps), whereas with $c_l = 7$ many more divergent cases remain, similar to the pure S case. Cases between $c_l = 3$ and $c_l = 5$ performed

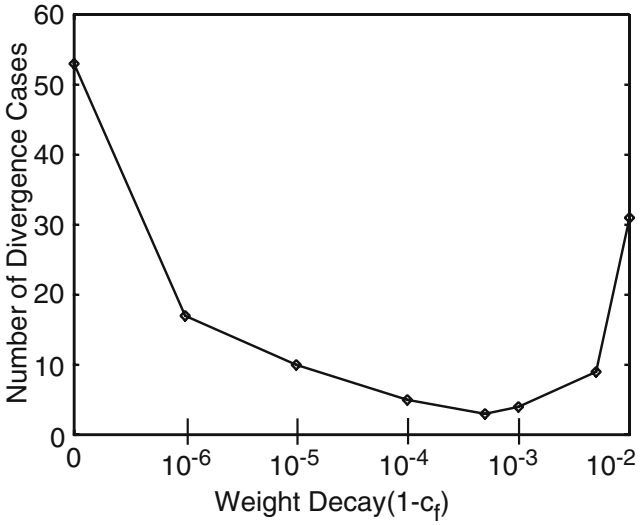

Fig. 6 Number of diverging cases for different degrees of weight decay using strategy SF

similarly, successfully diminishing the number of divergent cases. Exploration has similar effects for a wide range of values, from $p_e = 0.05$ to around 0.5. With increasing exploration, learning is mostly successful, but the paths get more disordered because of frequent off-path exploratory steps.

Concerning our real rats, time to convergence is more difficult to measure as animals always maintain a high degree of exploration drive and are easily distracted even when hungry. The following observations were made. Rats found for the first time a short and direct path to food after 6-10 trials (rat1: 9, rat2: 10, rat3: 6, rat4: 9, rat5: 7 trials). In rats 1 and 2 this was then followed by another exploratory phase and consistent food targeting was found for these two animals at around trial 20. Rats 3-5, on the other hand, continued to run to the food until not hungry anymore with some exploration around trial 15. Thus, convergence is faster than for the simulated rats which is probably due to the fact that the arena cannot be stripped of all visual and or self-generated odor cues (self-generated as a consequence of the rat's running), which both provide a very strong signal to target.

3.5 Place fields size and density

As discussed above, the convergence properties of RL-algorithms are not only affected by the path structure but also by the state space characteristics. This problem arises here as a consequence of the structure of our place field map. Thus, next we will address the question of how the obtained results depend on place cell radius and density. Hippocampal place fields often have a radius of around $1/4^{th}$-$1/5^{th}$ of the arena (10-20 cm), though smaller and bigger fields have also been observed, and the size may depends on the size of the arena (Wilson and McNaughton 1993; O'Keefe and Burgess 1996; Muller 1996; Mehta et al. 1997). In our model e.g. for σ = 4.24 cm we have a P=10%-firing field of 22 cm diameter within the 1.5 × 1.5 m. arena, and that matches well the observed size of place fields. Little is known about place field density, because from standard recordings in a single animal density is not straightforward to evaluate. Some authors find over-representations of places that are more important for an animal or are more densely explored (Hollup et al. 2001).

To investigate the influence of place field size and density we performed an exhaustive analysis over size/density pairings adjusted to lead to a similar coverage for the whole arena. Coverage in our model is calculated as the average number of cells that will actually fire at any given location. Coverage values of 0.8, 2.4 and 4.5 were investigated. Depending on their size, different numbers of cells were required for this, as given in the central part of Table 3.

Note, as coverage is not uniform, a certain part of the surface will always on average remain uncovered. For an average coverage of 4.5 we have about 1%, for 2.4 about 6% and for 0.8 about 45% of the surface area of the arena uncovered.

Convergence of paths was evaluated through path-length histograms for 100 trials each (Fig. 7). The best performing strategy (SEF) for path generation and learning was employed with default parameters. The shortest path found in these experiments contained 16 steps and path length was limited to 300 steps; a new trial was started if this number was exceeded (see x-axis labeling in panel d). Hence, in a given bin, we plot how many times the rat had found the reward within

Table 3 Number of cells required to achieve a certain coverage given the field width σ

Coverage	Field Width (σ) in cm			
	2.12	4.24	6.36	8.48
4.5	2000 (a)	500 (b)	230 (c)	140 (d)
2.4	1100 (e)	300 (f)	130 (g)	80 (h)
0.8	350 (i)	100 (j)	50 (k)	25 (l)

Value of $\sigma = 4.24$ cm, $n = 500$ have been used for most other experiments. Labels (a-l) refer to the panels in Fig. 7

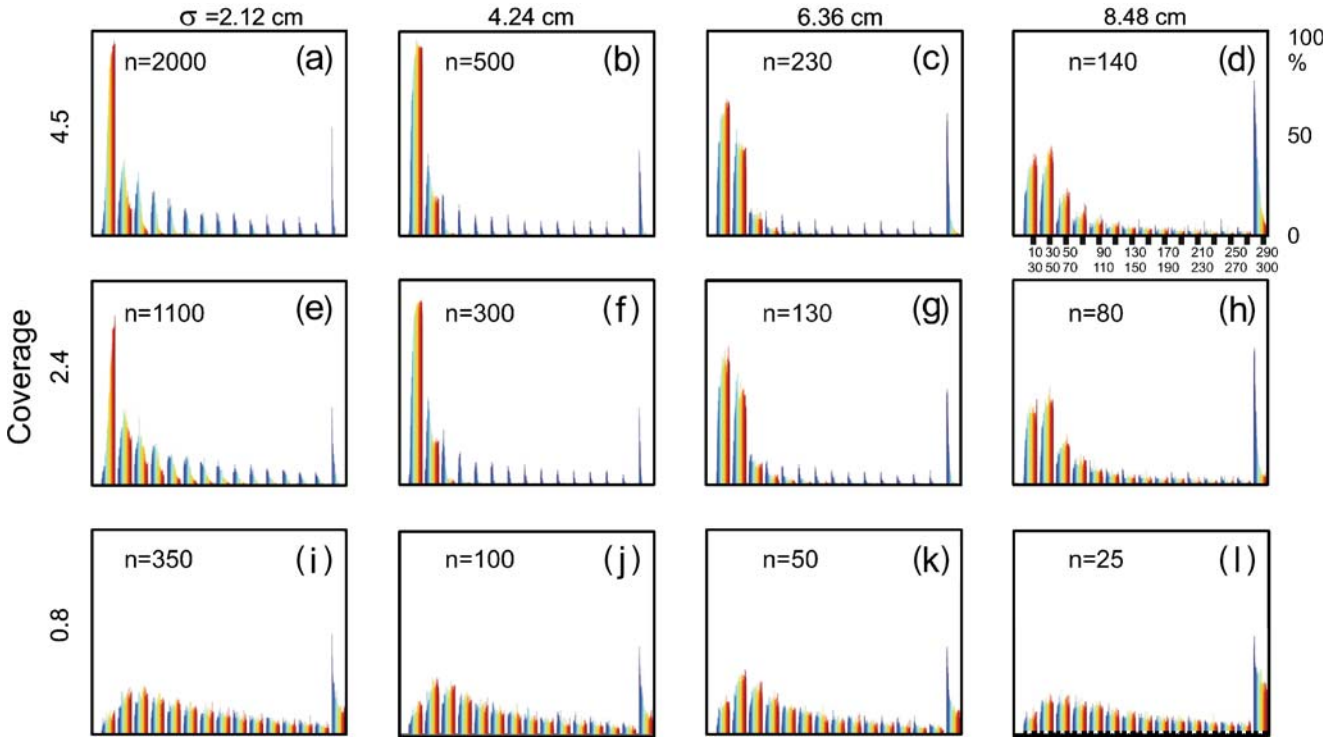

Fig. 7 Path length distributions for different combinations of place field size σ and number of cells n, leading to different degrees of coverage as shown on the left side (**a**–**l**). Colors encode the stage of learning (*blue*=beginning of learning, *red*=end). For further explanation see text

the number of steps with which that bin is labeled from 100 experiments performed. Clearly, early in learning, paths are longer and later they are, if convergent, shorter. To show this, histograms are color coded. Starting phase of learning is shown in blue (leftmost color in each bin), and ending phase in red (rightmost color in each bin). Middle stages are shown by the other colors. For each field size/density pair a separate histogram is provided in the figure. Histograms are given in the same order as in Table 3; they are normalized to 100.

Because 300 steps is the absolute path length limitation, which is often used in the first few trials, one finds a blue peak in all panels in the 300-bin. Next, we note that the peak of all blue contributions (early during learning) is, as expected, in all cases shifted to the right with respect to the red contributions (late during learning). Trivially, learning makes the paths shorter.

Looking in more detail at the red contributions (after learning), one can see that the shortest paths were found on average for panels (a) and (b), where the red peak falls in the 10-30 bin, where in (b) convergence is faster. In general many times the red peak falls in the 10-30 bin, but often the other colors are not much found in this bin. This means that the finally reached path length was indeed 10-30 steps but that this has been reached only after many (100-200) trials. In panel (b), on the other hand, also the blue color is represented with a high peak in the 10-30 bin, hence short paths have been already found early during learning. Panel (b) corresponds to $\sigma = 4.24$ cm, $n = 500$ and this combination has, therefore, been used for all experiments reported above. With smaller place fields $\sigma = 2.12$ cm (column one, a, e, i) convergence is much slower. Bigger place fields (right two columns) produce poorer convergence, introducing many red contributions into the higher order bins, pointing to long, final path lengths. Small coverage (last row), independent of the field width, produced poor convergence.

In summary one finds that high coverage corresponds to highly overlapping place fields, which in general increases the speed of convergence. For a given place field density, small place fields increase the proportion of convergent paths, but also increase the time to convergence. The effect of place field size could, thus, be understood to be the product of a trade-off between convergence and speed to convergence. This can be explained largely by the following observations: For bigger fields we get correlations from further away, which makes learning quicker, while for smaller fields,

we get a finer approximation of the space, which makes it more accurate and less prone to divergence.

4 Discussion

The goal of this study was to investigate how path formation strategies interact with RL in a place field-like system for action value function approximation. To this end we have first analyzed rat path characteristics in freely exploring rats showing that these paths contain a significant proportion of long straight segments. This analysis allowed us to model path formation more realistically also in the RL-simulation using what we called the S-strategy (path straightening). The RL-simulation was based on overlapping, probabilistically firing place fields. We found that the number of divergent cases with path straightening (S) increased two-fold compared to traditional exploration-exploitation strategies (E-strategy). Thus, increasing random exploration (E) improved convergence but this way we received unrealistic, wiggly paths. We showed that two biologically plausible mechanisms (weight decay and path length limitation) will improve the learning in such a (S or SE) system. Both mechanisms limit the memory of the system and thereby they act against trapping. Specifically we could show that convergence can be improved if weight decay (F) or path length limitation (L) are added to the path straightening setup. Briefly: For the S-strategy, weight decay recovers and improves performance while path length limitation reduces the divergence. Mixing too many strategies (e.g. using F *and* L) will in general not anymore lead to an improvement; rather often a deterioration of the performance is now found as both mechanisms act limiting on the memory of the system. Furthermore, we have demonstrated that the degree of improvement also relies on the size and degree of overlap in the place field system. Here, we showed that the effect of place field size and density will, essentially, lead to a trade-off between convergence and speed to convergence.

4.1 Relevance and influence of the chosen learning algorithm

In this study we have chosen SARSA learning for our experiments. This choice is motivated by several reasons. In the first place, SARSA is on-policy learning. Hence, an agent updates (synaptic) weights by the outcome of the actually performed action, different from the (more commonly used) Q-learning algorithm, where weights are being updated by the best *possible* action outcome, even if the agent has actually chosen a different *performed* action. Q-learning would thus require some kind or reasoning along the lines: "I know the best possible outcome and update my learning with this even though I am trying out something else (exploration)." It has been shown that such human-like "reasoning" does not seem to be represented in the midbrain dopaminergic system, which is the structure mostly held responsible for the implementation of reward based learning in the animal brain (Schultz 2002, 2007). At the level of monkeys it seems that SARSA learning prevails (Morris et al. 2006) as discussed by Niv et al. (2006). We have also performed quite an exhaustive analysis of Q-learning. In fact, as observed often for systems with function approximation (Tsitsiklis and Van Roy 1997; Sutton and Barto 1998; Wiering 2004), Q-learning performs far worse under a mixed (SE) strategy (data not shown). Paths will differ though. For example zigzagging observed in badly convergent SARSA is much less pronounced and will now be replaced by long curved paths in badly convergent Q-learning cases.[1] An interesting suggestion to possibly improve convergence of such systems would be to employ information from head-direction cells (for a review see Sharp et al. 2001) to augment the state space, which is currently represented by the place fields as such. To test this idea we have augmented the state-action space in a Q-learner by a head-direction system. We have performed a set of control experiments (data not shown) were path straightening had been implemented via a heading asymmetry, hence, through biasing Q-values for the different relative movement directions, including the same exploration tendency as in our SE-system. We observed that initial paths of this system are then indeed indistinguishable from those of the SE-system used here. However, when such a system learns, the first thing that happens is that the initial motion trajectories, which are still not targeting the reward, will destroy the directional bias reverting the system into one similar to our pure E-component. Hence, before convergence the simulated rat undergoes a phase where quite unrealistic, wiggly paths segments prevail. Furthermore, due to the larger state-action space, convergence is often slower. In view of this, we had decided against including the head-direction system into this model. This is also supported by the fact that there is conflicting evidence as to the influence of head-direction cells on rats' navigation behavior. Golob et al. (2001) showed

[1] Note, to be more specific, we have used SARSA($\lambda = 0$). There would also be the choice of using SARSA or Q with $\lambda \neq 0$. Speed of convergence in conventional RL can change as a consequence of λ. There are, however, in general no predictions possible for which value the fastest convergence is obtained.

that navigation cannot be predicted from head direction cell activity, while Dudchenko and Taube (1997) could show some influence on behavior.

Another alternative for further investigations would be to employ Actor-Critic modules (Barto et al. 1983; Barto 1995; Sutton and Barto 1998), for example using a temporal-difference (TD, Sutton 1988) based critic to decide about possible actions. Such a choice would be justifiable by the suggested relation of Actor-Critic architectures to the interfacing between basal ganglia, prefrontal cortex, and the motor system (Barto 1995; Houk et al. 1995). Indirect evidence exists that humans seem to be able to perform an off-line actor-critic update (O'Doherty et al. 2004). Indeed, a wide variety of such architectures has been suggested during the last years (Houk et al. 1995; Berns and Sejnowski 1998; Brown et al. 1999; Contreras-Vidal and Schultz 1999). However, their relation to the basal ganglia and prefrontal cortex remains rather abstract and these models are often quite incompatible with the physiology and anatomy of the biological substrate (discussed in great detail in Wörgötter and Porr 2005).

Actor-Critics usually rely on the interpretation of the dopaminergic signal as being the δ-error of TD learning, supported by studies of Schultz and collaborators (Schultz 2002, 2007; Contreras-Vidal and Schultz 1999; O'Doherty et al. 2004). More recently, the dopamine signal has been re-interpreted rather as a timing signal for the learning (Redgrave and Gurney 2006) as there appear to be timing conflicts between the different neuronal responses when using the traditional interpretation (Redgrave et al. 1999). This may also lead to a reinterpretation of the Actor-Critics idea and its relation to the neuronal circuitry.

On the more technical side it is known that the convergence of Actor-Critics is often quite difficult to achieve and there are many ways to construct such an architecture (Sutton and Barto 1998), which leaves the choice too unconstrained in conjunction with this investigation. Little is known about the behavior of Actor-Critics together with function approximation. As the possible relations of this different RL-algorithms to brain function is still a matter of debate (Wörgötter and Porr 2005, 2007), SARSA seems to be a justifiable choice.

4.2 Biological behavioral mechanisms

Our study is the first to consider rat path statistics during exploration in conjunction with navigation learning. Some other studies were concerned with open-field experiments and also there paths with long straight stretches have been observed (Etienne et al. 1996; Eilam 2004; Zadicario et al. 2005). In daylight animals tend to run along walls or hide in the corner, while at night some more exploration in the center of the field happens (Eilam 2004; Zadicario et al. 2005). Our path generation algorithms did not include all the complexity (loops, stopping) observed in actual rodent paths, but nevertheless our path geometry, and path statistics resemble real rat paths.

In real rats strategies similar to F or L are a common observation. Different forms of forgetting (weight decay) are common in animals and humans. The "frustration" (path length limitation) mechanism used here can be linked to the return-to-home-base drive in real rats. It is known that rats return to their home location in an open arena exploration process and investigate the environment in loops of increasing length (Eilam and Golani 1989; Whishaw et al. 2001; Wallace et al. 2002; Hines and Whishaw 2005; Zadicario et al. 2005; Nemati and Whishaw 2007). Furthermore there are again behavioral differences during daylight as compared to the night, where homing is less prevalent because animals are less frightened (Eilam 2004; Zadicario et al. 2005).

Our model makes relatively few assumptions about the determinants of behavior. We have not accounted for motivational state, thigmotaxic tendencies, arousal, and fatigue, to name but a few factors that likely influence path learning. To capture the full complexity of real animals, additional mechanisms would have to be considered. However, a strength of the current model is that with just a few assumptions we can produce a good approximation of actual rat navigation.

Some limitations may arise from the fact that start and goal locations for the real rats were not varied. However, as long as rats are not being directly attracted by other landmarks (e.g. starting them close to a wall), one should hope that path characteristics will not change too much. Observations in the water maze seem to support this notion and rats go after learning straight to the platform (Morris 1981; Foster et al. 2000), but quantitative data for these paths are not available. In the model, changing start and goal will only lead to a rotation of the Q-value vector field as there are no additional attraction or repulsion mechanisms built in.

There are some further considerations as to the contingencies of the animal experiments, which we would like to briefly discuss. For practical purposes animals found the reward always at a curtain in a wall. Orienting with respect to the walls (and their multiple curtains) represents a form of allothetic navigation, where the place field activity is driven by sensor inputs. Without explicitly having modeled such inputs, our Q-value system is also based on allothetic (coordinate based) information, which is the common way to model such

systems (see Recce and Harris 1996; Burgess et al. 2000; Foster et al. 2000 and for a discussion on idiothetic, path integration influences see e.g. Kulvicius 2008). The fact that the reward for the model rat has not been provided when touching a wall leads sometimes to more trials until convergence as compared to experiments with the reward given at a wall, because in the former case model rats can bypass the reward on the other side, too. In the model we have furthermore assumed that a minimal step of the model rat is 6 cm. This corresponds to about half the body length of a rat (without tail) and reflects the fact that we do not model sharp, local turns, where a rat would bend back on itself. This happens only when an animal stops, sniffs (most often), and turns, which we are not modeling as our model focuses only on continuous path segments. Furthermore, the step size is tuned to the place field size, making each step small enough for the model rat not to pass over place fields 'unnoticed' due to discretization effects.

4.3 Hippocampus modeling

It was not our objective to create a general hippocampus model (Kali and Dayan 2000; Becker 2005). We do not distinguish between regions of hippocampus (dentate gyrus, CA1, CA3), nor do we model its inputs or the process of place cell development (Samsonovich and McNaughton 1997; Tsodyks 1999; Hartley et al. 2000). Instead we focused on the interaction between behavioral constituents (paths) with RL in a place-field like function approximation system.

In the field of hippocampus-based navigation our model has similar properties to models by Arleo and Gerstner (2000), Krichmar et al. (2005), Strösslin et al. (2005). Those models try to incorporate many known details about the included brain structures and types of cells present, thus attempting to study hippocampal function. No special attention has been devoted in these studies to path formation and its influence on the learning. Implications how path characteristics could influence such studies will be discussed later.

4.4 Machine learning

We have emulated navigational learning using RL with function approximation, based on hippocampus-like place field representation. We used the SARSA algorithm to stay closer to biological learning mechanisms, though Q-learning can be implemented in the same framework, as discussed above. The algorithm is similar to the one proposed by Reynolds (2002). However, we do not normalize the learning equation and our learning rates are, thus, independent of the number of activated features (activated place fields). This is motivated by the problem of how to emulate such a global normalization in a neuronally plausible way. Global normalization by neuronal mechanisms is a well-known difficulty also for other simpler learning algorithms for example when wanting to limit weight growth in Hebbian learning (Dayan and Abbott 2005). In spite of the lacking normalization, the current algorithm produces convergent weights as well as convergent behavior in a conventional exploration-exploitation setup. On many occasions it produces optimal paths to reward. As our place field system fires probabilistically, it is difficult, if not impossible, to provide a rigorous convergence proof. In general, such proofs are notoriously hard to obtain for any function approximation system even under more relaxed conditions (Szepesvari and Smart 2004). While of possible theoretical interest, these machine learning related issues may not relate directly to our more biologically-inspired model.

In machine learning weight decay is known from the general purpose neural network learning literature as a means to prevent weights from saturation (Werbos 1988). In a RL framework, weight decay has been used in several isolated studies, usually to produce agents who can adapt to changing environments (Yen and Hickey 2004). Hence, as such the idea of using forgetting is not new, but here we show how learning in a static environment can also benefit from this mechanism in preventing divergence.

Path length limitation in our setup is implemented as return to the home base if the reward was not reached in predefined number of simulation steps. Path-to-goal limiting is a natural option for any simulation of RL; it is reasonable to stop a trial after some steps if the reward is not found (e.g. Glaubius and Smart 2004). Here we used a more complex path limiting process, where the allowed number of steps depends of the path-to-goal length in the previous epoch. In robotics applications with RL, path length limitation is often included to switch the pattern of behavior when the goal is not reached (e.g. Okhawa et al. 1998), which is not of relevance, though, for our system.

4.5 Possible relevance

The findings presented here could possibly influence at least three fields: biological modeling of place field based navigation, RL theory, and machine learning applications (robotics).

Several models for navigation exist based on hippocampal place fields (Arleo and Gerstner 2000; Krichmar et al. 2005; Strösslin et al. 2005; Sheynikhovich et al.

2005). In some of them eligibility traces are used (Arleo and Gerstner 2000; Strösslin et al. 2005; Sheynikhovich et al. 2005) for memorizing the most recent path segments. In our study, we could show that such a memory mechanism is not required to achieve efficient learning. It is unclear to what degree such eligibility traces exist. Thus, following a more conservative attitude one might choose our approach to simplify modeling.

In Strösslin et al. (2005) and Sheynikhovich et al. (2005) not only the action performed, but also actions from the spatially near states are involved in learning, thus introducing an averaging effect, that leads to better convergence in their case. SARSA or Actor-Critic learning do not foresee an action memory of this kind and such a neighborhood action excitation mechanism may be difficult to justify from a biological point of view. Alternatively, here we show that also other simpler mechanisms, weight decay and path length limitation, can improve convergence.

Hence in general the results presented here suggest that convergence in existing models could possible also be assured by mechanisms of weight decay and path length limitation thereby making it possible to eliminate aspects of unclear biological realism used so far.

Furthermore, none of the mentioned studies explicitly deals with path smoothing, and the examples in the study of Krichmar et al. (2005) clearly demonstrate that the paths are wiggly in the beginning. Their robot was very slow, stopping and orienting for navigation, quite unlike real animals. Thus, to assure that paths are realistic one needs to include mechanisms similar to the ones used here also in the other studies.

The field of RL theory is dominated by attempts to arrive at rigorous convergent proofs for their methods. However, several theoretically sound machine learning algorithms cannot be used without alterations in praxis as their proven convergence is far too slow in real world problems with large state and action spaces (temporal credit assignment problem, (Wörgötter and Porr 2007)). SARSA and Q-learning behave like this, too. Function approximation has been used to improve speed of convergence, but these systems are now many times very hard to analyse mathematically and - even worse - often not strictly convergent anymore. In this context it appears of interest that mechanism can indeed be found empirically (here F and L), which improve convergence while maintaining speed. Machine learning has increasingly started to investigate such "difficult" systems knowing that one needs both, speed and reliability (of convergence). The memory limiting aspect and the observed un-trapping when using mechanisms similar to F and L ought to be a general theoretical interest for RL-systems. It would, thus, be interesting, albeit probably quite difficult, to investigate such mechanisms from a mathematical point of view, using a more rigid state-action RL-system to better constrain the problem.

The aspects on RL theory discussed above directly reach out into the applied fields. It is worth noting that exploration-exploitation strategies as required by theory are in some cases totally incompatible with the compliance requirements of machines. Especially in multi-joint robot arms the very jerky, wiggly movements obtained by random exploration are not permissive as they will damage such a machine (T. Asfour, personal communication). Hence, straighter exploration paths should be employed for RL problems in these domains. The improved convergence found in our study could therefore help to better adapt RL-methods to such problems. Indeed, we have started to employ our methods now in the context of a humanoid robot (Asfour et al. 2006), learning to reach for a target currently still limited to 4 Degrees of Freedom in a 3-D reaching space. Preliminary results show that the machine can often learn this task using about 10000 "place fields" in only 20-30 trials taking a few minutes. In such a large state-action space, conventional machine learners without additional mechanisms would converge only after days, while their exploration patterns would damage the arm. Clearly there are other ways to implement efficient RL in such robot systems, but this example shows that our approach appears also promising.

Acknowledgements FW and PD acknowledge funding by Grant: BB/C516079/1 from the Biotechnology and Biological Sciences Research Council (U.K.); FW furthermore acknowledges funding by the European Commission "PACO-PLUS".

We thank Dr. D. Sheynikhovich and Prof. W. Gerstner for drawing our attention to the path finding problem.

Appendix: RL with function approximation

RL is a procedure where a value function $V(s)$ over states s develops as an agent acts in its environment and attains goals. In RL with delayed reward, the function shows a gradient towards a goal. In the Q and SARSA learning approaches (Watkins and Dayan 1992; Kaelbling et al. 1996), instead of the state value function, the state-action value function $Q(s, a)$ (short: *action value function*) is developed, where s denotes a

state and a an action. Action-value functions describe values of concurrent actions in every state, and can be directly used for making a decision on which action to perform.

Q-learning is described by the following equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right) \tag{5}$$

Where $Q(s_t, a_t)$ is the action value function at time step t, r_{t+1} is a reward obtained with action a_t, α is the learning rate and γ a discount factor. SARSA learning differs by a single aspect that the current action value is updated according to the value of the next *actual* action, but not by the *best possible* next action, as in Q learning:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha (r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \tag{6}$$

Hence, SARSA is designed to work on-policy, which means that learning takes place as an agent moves in the state space according to the path that was actually performed. Evidence exists that animals follow an on-policy, rather than an off-policy, learning strategy (Morris et al. 2006, see also commentary by Niv et al. 2006). Hence in this study we are investigating the SARSA algorithm and only sometimes comment on Q-learning.

For big and/or continuous state spaces, function approximation methods need to be used, where the action value function is a function of more abstract and wider-embracing entities commonly called *features* in the RL-literature. We define normalized Q-values by:

$$Q(s_t, a_t) = \sum_i \theta_{i,a_t} \Phi_i(s_t) / \sum_i \Phi_i(s_t) \tag{7}$$

where $\Phi_i(s_t)$ are the features over the state space, and θ_{i,a_t} are the adaptable weights binding features to actions (Reynolds 2002).

We assume that a place cell i produces spikes with a scaled Gaussian-shaped probability distribution:

$$p(\delta_i) = A e^{-(\delta_i^2 / 2\sigma^2)} \tag{8}$$

where δ_i is the distance from the i-th place field center to the sample point (x, y) on the trajectory, σ defines the width of the place field, and A is a scaling factor. In the areas where the values of this scaled distribution are above 1, cells fire with a probability of 1.

We then use the actual place field spiking to determine the values for features Φ_i, $i = 1, .., n$, which take the value of 1, if place cell i spikes at the given moment on the given point of the trajectory of the model animal, otherwise it is zero:

$$\Phi_i(s_t) = \begin{cases} 1 \text{ if place cell i spikes at } s_t \\ 0 \text{ else.} \end{cases} \tag{9}$$

SARSA learning then can be described by:

$$\theta_{i,a_t} \leftarrow \theta_{i,a_t} + \alpha (r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - \theta_{i,a_t}) \Phi_i(s_t) \tag{10}$$

where $\theta_{i,a}$ is the weight from the i-th place cell to action(-cell) a, and state s is defined by (x, y), which are the actual coordinates of the model animal in the field.

We sum over all features, but in each place only a specific subset of cells will fire rendering their corresponding features non-zero. Note that function $\Phi_i(s)$ has a probabilistic nature in our approach, differently from usual features used for function approximation in RL. The update rule Eq. (10) we use has a straightforward biological interpretation: the weight of a particular action is increased at the given place if this weight leads either to a reward, or if it leads on to pieces of an already known rewarding path. In the latter case this results from the non-zero $Q(s_{t+1}, a_{t+1})$-values in the next state.

References

Ainge, J. A., Tamosiunaite, M., Wörgötter, F., & Dudchenko, P. A. (2007). Hippocampal CA1 place cells encode intended destination on a maze with multiple choice points. *The Journal of Neuroscience*, *27*(36), 9769–9779.

Arleo, A., & Gerstner, W. (2000). Spatial cognition and neuromimetic navigation: A model of hippocampal place cell activity. *Biological Cybernetics*, *83*(3), 287–299.

Arleo, A., Smeraldi, F., & Gerstner, W. (2004). Cognitive navigation based on nonuniform Gabor space sampling, unsupervised growing networks, and reinforcement learning. *IEEE Transactions on Neural Networks*, *15*(3), 639–652.

Asfour, T., Regenstein, K., Azad, P., Schröder, J., Bierbaum, A., Vahrenkamp, N., et al. (2006). ARMAR-III: An integrated humanoid plattfrom for sensory-motor control. In: *In IEEE-RAS International Conference on Humanoid Robots*.

Barto, A. (1995). Adaptive critics and the basal ganglia. In: J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge: MIT.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*, 835–846.

Becker, S. (2005). A computational principle for hippocampal learning and neurogenesis. *Hippocampus*, *15*(6), 722–738.

Berns, G. S., & Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *The Journal of Cognitive Neuroscience*, *10*(1), 108–121.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *The Journal of Neuroscience*, *19*(23), 10502–10511.

Brown, M. A., & Sharp, P. E. (1995). Simulation of spatial learning in the Morris water maze by a neural network model of the hippocampal formation and nucleus accumbens. *Hippocampus*, *5*(3), 171–188.

Burgess, N., Jackson, A., Hartley, T., & O'Keefe, J. (2000). Predictions derived from modelling the hippocampal role in navigation. *Biological Cybernetics*, *83*, 301–312.

Contreras-Vidal, J. L., & Schultz, W. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of Computational Neuroscience*, *6*, 191–214.

Dayan, P., & Abbott, L. F. (2005). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge: MIT.

Dudchenko, P. A., & Taube, J. S. (1997). Correlation between head-direction single unit activity and spatial behavior on a radial arm maze. *Behavioral Neuroscience*, *111*, 3–19.

Eilam, D. (2004). Locomotor activity in common spiny mice (Acomys cahirinuse): The effect of light and environmental complexity. *BMC Ecology*, *4*(16), 4–16.

Eilam, D., & Golani, I. (1989). Home base behavior of rats (rattus norvegicus) exploring a novel environment. *Behavioural Brain Research*, *34*, 199–211.

Etienne, A. S., Maurer, R., & Seguinot, V. (1996). Path integration in mammals and its interaction with visual landmarks. *Journal of Experimental Biology*, *199*(Pt 1), 201–209.

Foster, D. J., Morris, R. G., & Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, *10*(1), 1–16.

Glaubius, R., & Smart, W. D. (2004). Manifold representations for value function approximation. In: *Proceedings of the AAAI-04 workshop on learning and planning in markov processes* (pp. 13–189).

Golob, E. J., Stackman, R. W., Wong, A. C., & Taube, J. S. (2001). On the behavioural significance of head direction cells: Neural and behavioral dynamics during a spatial memory task. *Behavioral Neuroscience*, *115*, 285–304.

Hartley, T., Burgess, N., Lever, C., Cacucci, F., & O'Keefe, J. (2000). Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus*, *10*(4), 369–379.

Hines, D. J., & Whishaw, I. Q. (2005). Home bases formed to visual cues but not to self-movement (dead reckoning) cues in exploring hippocampectomized rats. *European Journal of Neuroscience*, *22*, 2363–2375.

Hollup, S. A., Molden, S., Donnett, J. G., Moser, M. B., & Moser, E. I. (2001). Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *The Journal of Neuroscience*, *21*(5), 1635–1644.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge: MIT.

Kaelbling, L. P., Littman, M., & Moore, A. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kali, S., & Dayan, P. (2000). The involvement of recurrent connections in area CA3 in establishing the properties of place fields: A model. *The Journal of Neuroscience*, *20*, 7463–7477.

Krichmar, J. L., Seth, A. K., Nitz, D. A., Fleischer, J. G., & Edelman, G. M. (2005). Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions. *Neuroinformatics*, *3*(3), 197–221.

Kulvicius, T., Tamosiunaite, M., Ainge, J., Dudchenko, P., & Wörgötter, F. (2008). Odor supported place cell model and goal navigation in rodents. *Journal of Computational Neuroscience*. doi:10.1007/s10827-008-0090-x.

Mehta, M. R., Barnes, C. A., & McNaughton, B. (1997). Experience-dependent, asymmetric expansion of hippocampal place fields. *Proceedings of the Natianal Academy of Sciences*, *94*, 8918–8921.

Morris, R. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation*, *12*, 239–260.

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, *11*(1), 47–60.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, *9*(8), 1057–1063.

Muller, R. (1996). A quarter of a century of place cells. *Neuron*, *17*, 813–822.

Nemati, F., & Whishaw, I. Q. (2007). The point of entry contributes to the organization of exploratory behaviour of rats on an open field: An example of spontaneous episodic memory. *Behavioural Brain Research*, *182*, 119–128.

Niv, Y., Daw, N. D., & Dayan, P. (2006). Choice values. *Nature Neuroscience*, *9*(8), 987–988.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*, 452–454.

O'Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, *381*(6581), 425–428.

Okhawa, K., Shibata, T., & Tanie, K. (1998). Method for generating of global cooperation based on local communication. In: *Proceedings of the 1998 IEEE/RSJ intl. conference on intelligent robots and systems* (pp. 108–113).

Recce, M., & Harris, K. D. (1996). Memory for places: A navigational model in support of Marr's theory of hippocampal function. *Hippocampus*, *6*(6), 735–748.

Redgrave, P., & Gurney, K. N. (2006). The short-latency dopamine signal: A role in discovering. *Nature Reviews Neuroscience*, *7*(12), 967–975.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends in Neurosciences*, *22*(4), 146–151.

Reynolds, S. I. (2002). The stability of general discounted reinforcement learning with linear function approximation. In: *UK workshop on computational intelligence (UKCI-02)* (pp. 139–146).

Samsonovich, A., & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *The Journal of Neuroscience*, *17*(15), 5900–5920.

Save, E., Nerad, L., & Poucet, B. (2000). Contribution of multiple sensory information to place field stability in hippocampal place cells. *Hippocampus*, *10*(1), 64–76.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*, 241–263.

Schultz, W. (2007). *Reward Signals.* Scholarpedia, http://www.scholarpedia.org/article/Reward_Signals.

Sharp, P. E., Blair, H. T., & Cho, J. (2001). The anatomical and computational basis of the rat head-direction cell signal. *Trends in Neurosciences*, *24*(5), 289–94.

Sheynikhovich, D., Chavarriaga, R., Strösslin, T., & Gerstner, W. (2005). Spatial representation and navigation in a bio-inspired robot. In: S. Wermter (Ed.), *Biomimetic Neural Learning for Intelligent Robots: Intelligent Systems, Cognitive Robotics, and Neuroscience* (pp. 245–264). New York: Springer.

Strösslin, T., Sheynikhovich, D., Chavarriaga, R., & Gerstner, W. (2005). Robust self-localisation and navigation based on hippocampal place cells. *Neural Networks*, *18*(9), 1125–1140.

Stuart, A., Ord, K., & Arnold, S. (1999). *Kendall's advanced theory of statistics.* London: Arnold, a member of the Hodder Headline Group.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9–44.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction.* Cambridge: MIT.

Szepesvari, C., & Smart, W. D. (2004). Interpolation-based Q-learning. In: *Twenty-First international conference on machine learning (ICML04)* (vol. 21, pp. 791–798).

Tesauro, G. (1995). Temporal difference learning and TD-gammon. *Communications of the ACM*, *38*(3), 58–67.

Tsitsiklis, J. N., & Van Roy, B. (1997). An Analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, *42*(5), 674–690.

Tsodyks, M. (1999). Attractor neural network models of spatial maps in hippocampus. *Hippocampus*, *9*(4), 481–489.

Wallace, D. G., Gorny, B., & Whishaw, I. Q. (2002). Rats can track odors, other rats, and themselves: Implications for the study of spatial behavior. *Behavioural Brain Research*, *131*(1–2), 185–192.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292.

Werbos, P. J. (1988). Backpropagation: Past and future. In: *Proceedings of the IEEE international conference on neural networks* (pp. 343–353).

Whishaw, I. Q., Hines, D. J., & Wallace, D. G. (2001). Dead reckoning (path integration) requires the hippocampal formation: Evidence from spontaneous exploration and spatial learning tasks in light (allothetic) and dark (idiothetic) tests. *Behavioural Brain Research*, *127*(1–2), 49–69.

Wiering, M. (2004). Convergence and divergence in standard averaging reinforcement learning. In: J. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Proceedings of the 15th European conference on machine learning ECML'04* (pp. 477–488).

Wilson, M. A., & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, *261*(5124), 1055–1058.

Wörgötter, F., & Porr, B. (2005). Temporal sequence learning, prediction, and control: A review of different models and their relation to biological mechanisms. *Neural Computation*, *17*(2), 245–319.

Wörgötter, F., & Porr, B. (2007). *Reinforcement Learning.* Scholarpedia, http://www.scholarpedia.org/article/Reinforcement_Learning.

Yen, G. G., & Hickey, T. (2004). Reinforcement learning algorithms for robotic navigation in dynamic environments. *ISA Transactions*, *43*(2), 217–230.

Zadicario, P., Avni, R., Zadicario, E., & Eilam, D. (2005). 'Looping': An exploration and navigation mechanism in a dark open field. *Behavioural Brain Research*, *159*, 27–36.